Culture Conflicts in Software Engineering Technology Transfer

Marvin V. Zelkowitz[*]
Department of Computer Science
and
Inst. for Advanced Computer Studies
University of Maryland
College Park, Maryland 20742
and Fraunhofer Center - Maryland
College Park, Maryland 20742

Dolores R. Wallace
Information Technology Laboratory

Natl. Inst. of Standards and Technology
Gaithersburg, Maryland 20899

David W. Binkley
Computer Science Department

Loyola College
Baltimore, Maryland
and Information Technology Lab.
Natl. Inst. of Standards and Technology
Gaithersburg, MD 20899

Abstract

Although the need to transition new technology to improve the process of developing quality software products is well understood, the computer software industry has done a poor job of carrying out that need. All too often new software technology is touted as the next "silver bullet" to be adopted, only to fail and disappear within a very short period. New technologies are often adopted without any convincing evidence that they will be effective, yet other technologies are ignored despite the published data that they will be useful. Clearly there is a clash between those developing new technologies and those responsible for developing quality products. In this paper we discuss a study conducted among a large group of computer software professionals in order to understand what techniques can be used to support the introduction of new technologies, and to understand the biases and opinions of those charged with researching, developing or implementing those new technologies. This study indicates which evaluation techniques are viewed as most successful under various conditions. We show that the research and industrial communities do indeed have different perspectives, which leads to a clash between the goals of the technology researchers and the needs of the technology users.

Keywords: Experimentation, Survey, Technology transfer, Validation models

1. Introduction

When the computer industry began several decades ago, software engineering was somewhat unique among engineering fields in that researchers and practitioners worked closely together in using and understanding this new technology. There was easy cross-fertilization between these two communities. Over time, this has changed with tremendous growth of computer applications, computer users, and computing professionals. Programming languages have evolved from-low level assembler languages to today's very high level visual object-oriented languages. Simple programs have become complex large systems, with some systems running an entire enterprise. Methods for developing programs have grown from design-writing on napkins to a myriad of overlapping processes comprising varieties of methods and documentation types.

[*] Research supported in part by National Science Foundation grant CCR-9706151 to the University of Maryland.

A response to this growth has been a corresponding growth in organizations dedicated to supplying an ever-increasing need for better tools and techniques for producing these complex products. Trade shows, research conferences, trade magazines proliferate on the technology scene. New professional technical journals regularly come alive to add to an already large number; the IEEE alone through its Computer Society currently publishes 20 monthly or bimonthly computer technology publications.

In spite of an abundance of methods and tools and information about them, why do the same problems appear over and over again in new software developments? Why are development schedules not met? Why do some systems fail? Why do some technical problems remain unsolved? While new solutions are frequently proposed, many have not been transferred into the industry at large. Many problems remain untouched by researchers. Why does it appear that today researchers and practitioners are no longer necessarily understanding each other's needs and efforts?

Researchers have been looking at the role of experimentation in computer science research [Fenton94]. However, most of these have looked at the relatively narrow scope of how to conduct replicated scientific experiments within this domain. We have been looking at the larger problems of the role of experimentation as an agent in transferring new technology into industry. We have been studying various experimental methods, in addition to the replicated experiment, useful for validating newly developed software technology [Zelkowitz97] [Zelkowitz98], and we have also studied various evaluation methods industry uses before adopting a new technology. As we later explain, these two processes are very different. The questions important to us include "Which of these validation and evaluation methods are most effective?" "Why aren't these methods used more often?" and "Why don't these results provide evidence for the transference of a technology into industry?" To try to understand these questions, we decided to survey a cross section of computer professionals about their views about software engineering technology validation.

1.1 The research and industrial communities

Researchers, whether in academia or industry, have a desire to develop new concepts and are rewarded when they produce new designs, algorithms, theorems, and models. The "work product" in this case is often a published paper demonstrating the value of their new technology. Development professionals, however, have a desire and are paid to produce a product using whatever technology seems appropriate for the problem at hand. The end result is a product that produces revenue for their employer.

Researchers select their research according to a topic of their own interest; the topic may or may not be directly related to a specific problem faced by industry. After achieving a result that they consider interesting, they have a great desire to get that result in print. Providing a good scientific validation of the technology is often not necessary for publication, and several studies have shown that experimental validation of computer technology is particularly weak, e.g., [Tichy95] [Zelkowitz98].

In industry, producing a product is most important and the "elegance" of the process used to produce that product is less important than achieving a quality product on time as a result. Being "state of the art" in industry often means doing things as well (or as poorly) as the competition, so there is considerable risk aversion to try a new technology unless the competition is also using it.

Consequently, researchers produce papers outlining the values of new technology, yet industry often ignores that advice. Assorted "silver bullets" are proposed as solutions to the "software crisis" without any good justification that they may be effective, are used for a time by large segments of the community, and

then are discarded when they indeed turn out not to be *the* solution. Clearly the research community is not generating results that are in tune with what industry needs to hear, and industry is making decisions without the benefit of good scientific developments. The two communities are severely out of touch with one another. The purpose of our survey is to try and understand these communities and understand their differences.

1.2 Research models

We began our effort to understand the differences between the research and industrial communities by examining models of experimentation for computer technology research. We identified 12 methods of experimentation that have been used in the computer field [Table 1.1] and verified their usage by studying 612 papers appearing in three professional publications at 5-year intervals [Zelkowitz98] from 1985 through 1995. About 20% of the papers contained no validation at all and another third contained only a weak ineffective form of validation. The figure for other scientific fields was more like 10% - 15% [Zelkowitz97]. The methods are defined in Appendix 1.

Table 1.1 Experimental Validation Models	
Case study	Project monitoring
Dynamic analysis	Replicated
Field study	Simulation
Legacy data	Static analysis
Lessons learned	Synthetic
Literature search	Theoretical analysis

Our results were consistent with those found by Tichy in his 1995 study of 400 research papers [Tichy95]. He found that over 50% of the design papers did not have any validation in them. In a more recent paper [Tichy98], Tichy makes a strong argument that more experimentation is needed and refutes several myths deprecating the value of experimentation.

1.3 Transition models

Given the set of research validation methods, we then sought to determine the techniques actually used by industry in order to transition a new technology. We visited several large development corporations[1] and interviewed reasonably high level individuals, such as Chief Scientist, Chief Technology Officer, and managers of large divisions. All had ultimate responsibility for technology selection. They were primarily influenced by trade shows, weekly trade magazines, Web information, customer opinion (i.e., technologies that would win the contract), vendor opinion, friends in other companies, and infrequently by the papers in professional technical journals. Sometimes recommendations from technical staff would be based on their readings and would eventually reach the managers' offices. Once a technology was identified, the companies might perform a pilot study or were mentored by an expert of the technology to determine if the technology would be effective.

Based on these industrial interviews and some earlier work by Brown and Wallnau [Brown96], we defined a set of industrial transition models for technology evaluation. While the transition models include some that are similar to those of the researchers, many are different [Table 1.2]; Appendix 2 provides a short description of these models. For example, vendor opinion (e.g., trade shows, weekly trade magazines, web

[1] To assure frank discussion, we agreed not to reveal the names of the corporations who spoke with us.

information) seemed important to industry; Web information also provides access to research literature so we needed to separate the medium in which information is located from the type of model that information supports. An important finding, though, is that everyone with whom we spoke claimed to use the web to find technology information.

Table 1.2 Industrial Transition Models	
Case study	Research literature
Data mining	Shadow (replicated) project
Demonstrator projects	State of the art
Feature benchmark	Survey
Field study	Theoretical analysis
Measurement	Vendor opinion
Pilot study	

1.4 Understanding each community

Researchers principally use methods from Table 1.1 in order to demonstrate the value of their technological improvements and industry selects new technology to employ by using the methods in Table 1.2. How do these communities interact? How can their methods support forward growth in computer technology and its application in real systems? We need to develop a better understanding of what each community understands and values. Then, perhaps, we can identify commonalities and gaps, and from there, mechanisms to enable each community to benefit better from the other.

2. Development of the survey

To understand the different perceptions between those who develop technology and those who use technology, we decided to survey the software development community to learn how they view the effectiveness of the various evaluation models of Tables 1.1 and 1.2. For questions, we based our survey on a previous survey [Daly97], modified for our current purposes. Each survey participant was to rank the difficulty of each of our 12 experimental models (or 13 evaluation models) according to 7 criteria, criteria 1 and 2 being new and 3 through 7 being the same as the Daly criteria. We decided to try to obtain an objective score by having all values ranked between 1 and 20, with 10 being arbitrarily defined as the maximum difficulty that a given company would apply in practice, and 20 being defined as an impossible model for that criterion.

2.1 Survey questions

The 7 questions we chose were:

1. *How easy is it to use this method in practice?* -- Can we use this method to evaluate a new technology? The answer should be independent of whether the method gives accurate results or not.
2. *What is the cost of adding one extra subject to the study?* -- Assume you want to add an additional subject (another data point) to your sample. What is the relative cost of doing so?

3. *What is the internal validity of the method?* -- What is the extent to which one can draw correct causal conclusions from the study? That is, to what extent can the observed results be shown to be caused by the manipulated dependent experimental variables and not by some other unobserved factor?
4. *What is the external validity of the method?* -- What is the extent to which the results of the research can be generalized to the population under study and to other settings (e.g., professional programmers, organizations, real projects)?
5. *What is the ease of replication?* -- What is the ease with which the same experimental conditions can be replicated (internally or externally) in subsequent studies? It is assumed that the variables that can be controlled (i.e., the dependent variables) are to be given the same value.
6. *What is the potential for theory generation?* -- What is the potential of the study to lead to unanticipated a priori and new causal theories explaining a phenomenon? For example, exploratory studies tend to have a high potential for theory generation.
7. *What is the potential for theory confirmation?* -- What is the potential of the study to test an a priori well defined theory and provide strong evidence to support it?

In an eighth question we asked each participant to rank the relative importance (again using the 1-20 ranking) of each of the 7 questions when making a decision on using a new technology. That is, which of the 7 questions was most important when a new technology was being evaluated?

We developed two different survey instruments from these 8 questions -- one by ranking each of the 12 research validation methods of Table 1.1 (i.e., the research survey) and one by ranking each of the 13 evaluation methods of Table 1.2 (i.e., the industrial survey).

2.2 Population samples

For our 2 survey instruments we obtained three populations to sample. Sample 1 included U.S.-based authors with email addresses published in several recent software engineering conference proceedings[2]. These were mostly research professionals, although included a few developers. Approximately 150 invitations to participate were sent to these individuals, and 45 accepted. The survey was not sent until the participant agreed to fill out the form, which we estimated would take about an hour to 90 minutes to read and fill out. About half of the individuals returned the completed form.

Sample 2 included U.S.-based authors with email addresses from several recent industry-oriented conferences. They were sent the industrial survey. About 150 invitations to participate were sent and about 50 responded favorably to our invitation. They were then sent the survey. Again, about half completed and returned the form.

Sample 3 were students in a graduate software engineering course at the University of Maryland taught by one of the authors of this paper. This sample was given the research survey. This course was part of a masters degree program in software engineering, and almost all of the students were working professionals with experience ranging up to 24 years. Not surprisingly, the return rate of the form for this sample was high at 96% (44 of 46).

It is important to realize that we wanted the subjective opinion of those surveyed on the value of the respective validation techniques based upon several criteria. Not everyone returning the survey had

[2] The survey was conducted via email.

previously used all, or even any, of the listed methods. We simply wanted their views on how important they thought the methods were. However, by choosing our sample populations from those writing papers for conferences or taking courses for career advancement, we believe we have chosen sample populations that are more knowledgeable, in general, about validation methods than the average software development professional. The invitations were sent early in 1998, and data was collected February through early April, 1998. Table 2.1 summarizes the 3 sample populations.

Table 2.1 Characteristics of each survey sample							
Sample	**Survey**	**Sample size**	**Years exper.**	**Academic Position**	**Industrial R&D**	**Industrial developer**	**Other (e.g., Consultants)**
1 (Research)	Research	18	18.6	9	3	3	3
2 (Industry)	Industry	25	19.1	0	5	8	12
3 (Students)	Research	44	6.6	1	5	27	11

3 Survey results

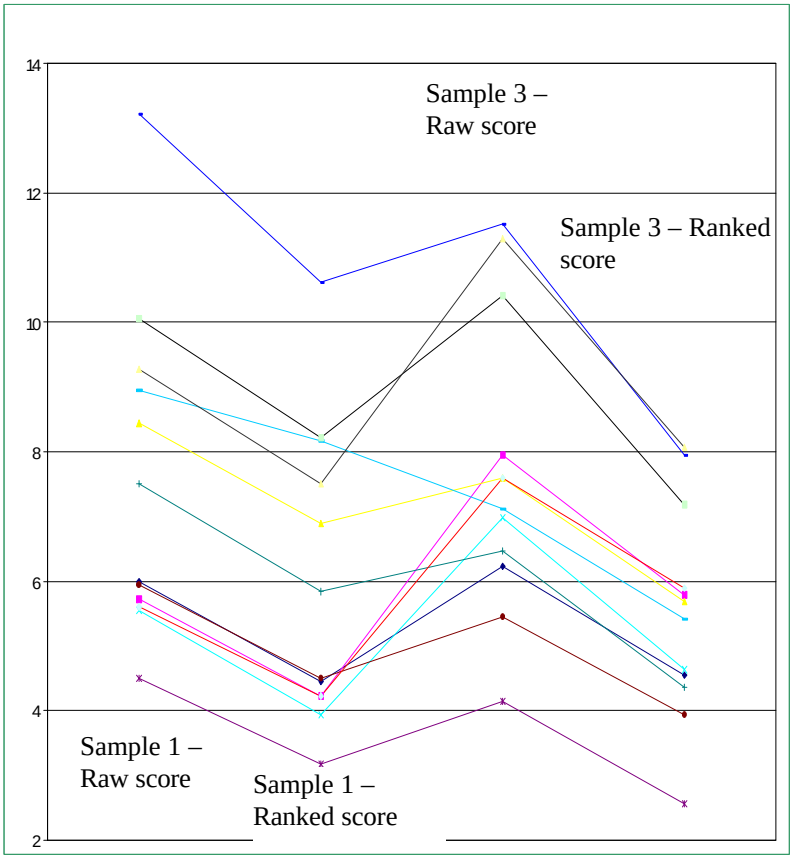


Figure 1. Easy to do. Average value for each of 12 validation methods.

Our initial concern was to determine bias in the set of responses. Would certain individuals rank all techniques high or low compared to other individuals? In order to test for this, we computed the average raw scores for each technique for each question, and we also ranked each answer (i.e., computing the easiest technique for each question, second easiest, third easiest, …, 12th easiest). This would eliminate such bias, but would also eliminate the significance of the value 10 being the subjective value of "hard to do." Fortunately, we believe that we don't have to take this into account. Figure 1 shows the value for the question "Easy to do." The first column represents the average raw scores for the 12 methods of Table 1.1 from the research sample (sample 1) and the second column is the average ranked score. Low values indicate the more important techniques. The fact that the ordering of the techniques from best to worst was essentially the same indicates that the raw score is an accurate reflection of the ranking. Only the 3rd and

4th, 5th and 6th, and 9th and 10th techniques switched places, not a major change. Columns 3 and 4 represent similar data from the student sample (sample 3). Here only the 3rd and 4th and 8th and 9th techniques switched places. However, there are some slight differences between sample 1 and sample 3, which will be discussed in Section 4.

Similar charts were obtained from the other questions. In addition, the correlation between the raw scores and the ranked scores for sample 1 was .86, .96 for sample 2 and .93 for sample 3. On this basis, we decided we could use the raw data and did not need to use only the ranked data for comparisons.

The average value for each technique for each of the 7 criteria appears in Figures 2 through 4. Figure 2 represents the average score for each of the 12 experimental methods over all 7 criteria for sample 1 with alpha=.05 confidence interval bars surrounding each average value. The "7" in each criterion represents the midpoint among the methods in order to make it easier to read the figure. Of greatest interest are bars that do not overlap, meaning there is a 95% probability that the average values for those techniques indicate a significant difference. Figure 3 represents a similar graph for sample 2 (the industrial group ranking 13 techniques) and Figure 4 represents a similar graph for sample 3 (the student industrial sample).

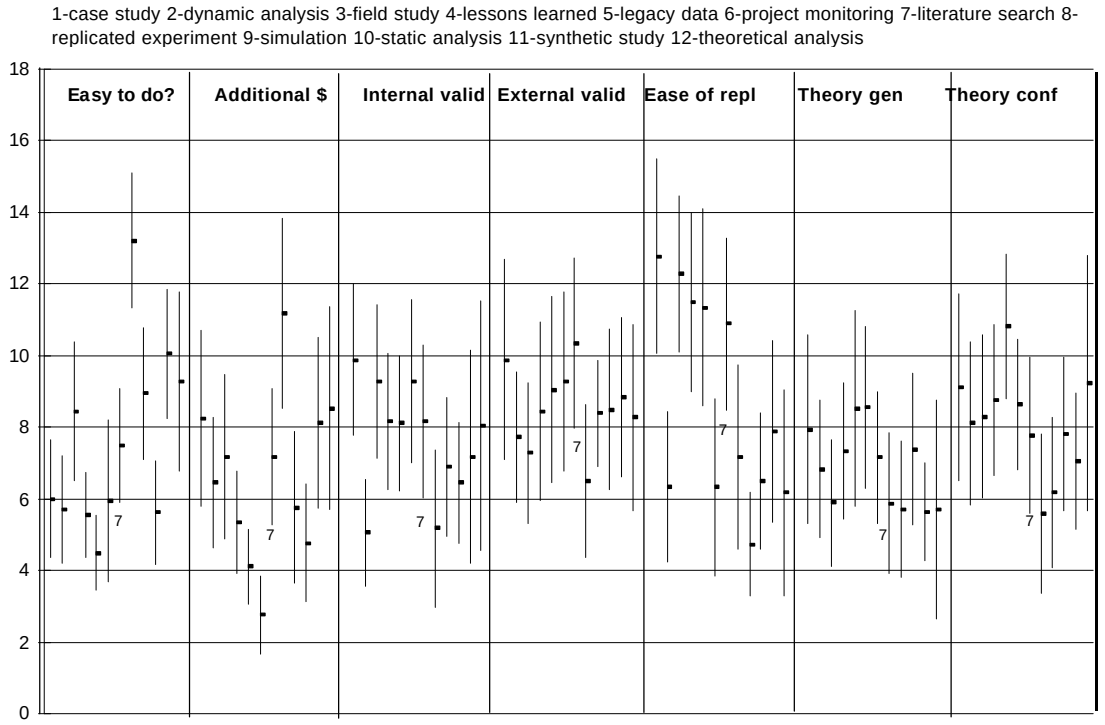


Figure 2. Sample 1 (research group) results.

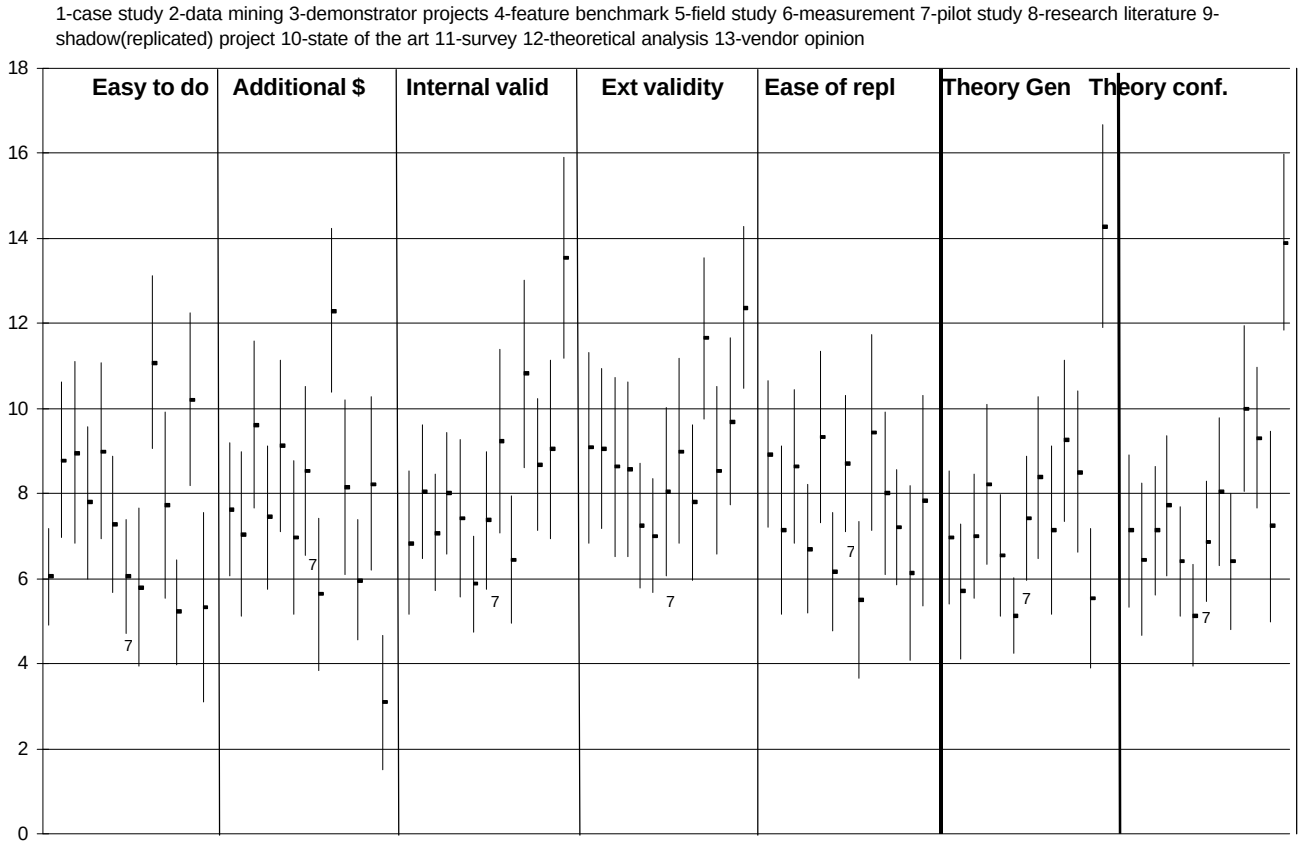


Figure 3. Sample 2 (industry group) results.

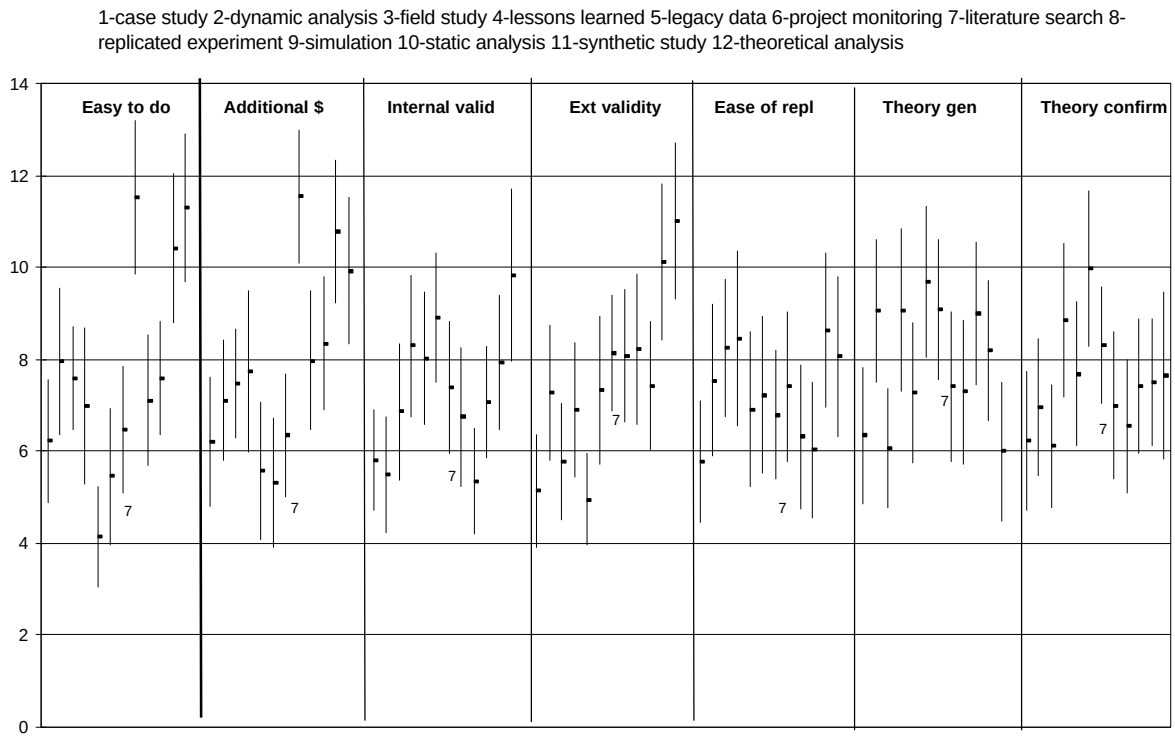


Figure 4. Sample 3 (student industrial group) results.

One way to simplify the data from these figures is to split the methods for each criterion into three partitions: practical, neutral, and impractical. The following procedure was applied:

1. Each method whose upper confidence interval was below the average value for all techniques would be listed in the practical partition. These methods are all "better than average" according to our 95% confidence criterion.
2. Each method whose lower confidence interval was above the average value for all methods would be listed in the impractical partition. These methods are all "worse than average" according to our 95% confidence criterion.
3. All other methods would be listed in the neutral partition.

Tables 3.1 through 3.3 summarize this process giving the practical and impractical techniques. All other methods are in the neutral partition.

Table 3.1 Practical and impractical techniques from research sample							
	Easy	**Addit. $**	**Int. val.**	**Ext. val.**	**Ease of repl.**	**Theory gen.**	**Theory conf.**
Practical	Dyn. anal Les. learned Legacy data Static anal.	Legacy data Proj. mon. Static anal.	Dyn. anal. Replication		Dyn. anal. Simulation Static anal.		Replicated
Impractical	Replicated Synthetic	Replicated	Case study		Case study Field study Les. learned		Legacy data

Table 3.2 Practical and impractical techniques from industry sample							
	Easy	**Addit. $**	**Int. val.**	**Ext. val.**	**Ease repl.**	**Theory gen.**	**Theory conf.**
Practical	Case study Pilot study Survey Vendor opin.	Res. Lit Survey Vendor opin.	Measure	Field study Measure	Measure Res. Lit.	Data mining Measure Theory anal.	Field study Measure
Impractical	Replicated	Replicated	State of art Vendor opin	State of art Vendor opin		Vendor opin.	State of art Vendor opin

Table 3.3 Practical and impractical techniques from student industrial sample							
	Easy	**Addit. $**	**Int. val.**	**Ext. val.**	**Ease repl.**	**Theory gen.**	**Theory conf.**
Practical	Case study Legacy data Proj. mon.	Case study Legacy data Proj. mon. Lit. search	Case study Dyn. anal. Simulation	Case study Legacy data	Case study	Case study Field study Theory anal.	Field study
Impractical	Replicated Synthetic Theory anal.	Replication Synthetic Theory anal.	Proj. mon. Theory anal.	Synthetic Theory anal.		Proj. mon.	Proj. mon.

Our final 8th question was to rate the importance of each of the 7 questions when making a decision on using a new technology. The purpose was to determine which of the criteria was most important when making such a decision. Figure 5 summarizes those answers on a single chart, the column labeled 1 representing the average values for the first sample, column 2 representing the average value for sample 2 and column 3 being sample 3.

4 Survey Evaluation

4.1 Preferred research techniques

Figures 2 and 4 and Tables 3.1 and 3.3 present a summary of our findings for the research validation methods. We summarize some of the observations from those figures.

In terms of easiness (question 1), replicated experiments and synthetic experiments for the research sample and replicated experiments, synthetic experiments and theoretical analysis for the student industrial sample were viewed as significantly (at the .05 level) harder to do than the other techniques and as impractical according to Tables 3.1 and 3.3. With average scores above 10, the consensus of these groups was that industry would never use such techniques as part of a validation strategy. It is no wonder that such techniques are rarely reported in the literature. In our earlier survey [Zelkowitz98] only 3.2% of the reported studies used synthetic or replicated experiments.

On the other hand, these two groups differed in their belief in the effectiveness of theoretical analysis with respect to internal and external validity (questions 3 and 4). Whereas the research group considered a theoretical validation likely to be used as much as any other technique (i.e., in the neutral partition of Table

3.1), the industrial group considered it most difficult to use, preferring instead the "hands on" techniques over the more formal arguments.

Other than the cost and ease issues, none of the other criteria exhibited significant differences among the respondents. However, when we combine the criteria into a single composite number, differences do become apparent (See Section 4.3).

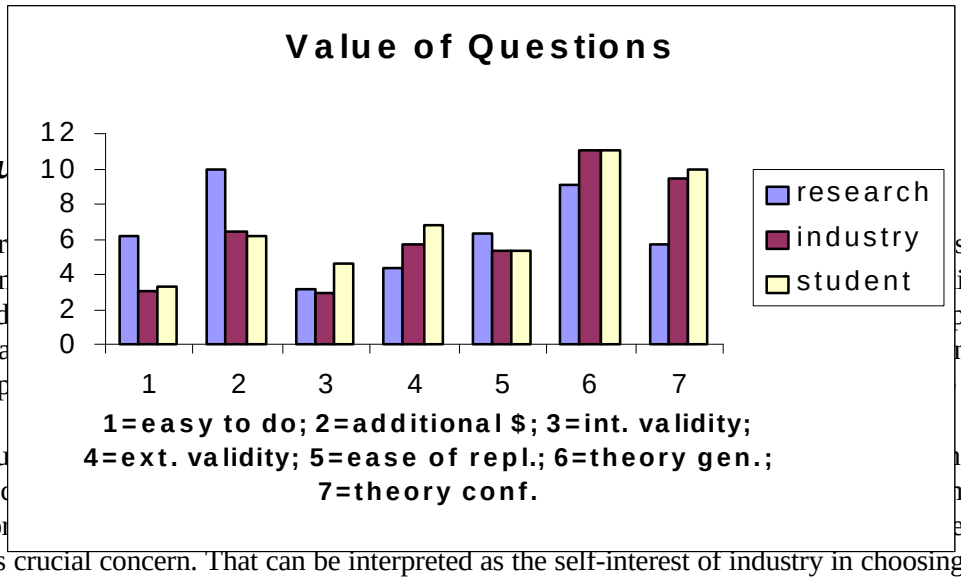
4.2 Preferred industrial methods

Figure 3 and table 3.2 give the basic results for the industrial transition methods. As with the research population, the replicated (shadow) project had an average rating (over all 7 questions) of over 10, signifying little industrial interest in performing such studies. Vendor opinion also averaged above 10, as did the need to be state of the art.

These high scores were all probably due to different reasons. Replicated experiments were viewed as hardest to do (highest score among all techniques at about 13.5), while vendor opinion had the worst internal and external validity (the ability for the method to explain the phenomenon under study, i.e., trusting the vendor to give the correct explanation). On the other hand, the need to be state of the art also suffered with respect to internal and external validity.

It is interesting to note that according to table 3.2, vendor opinion was considered practical according to ease of use (criterion 1), yet was impractical according to the criteria that dealt with accuracy of the evaluation (questions 3, 4, 6 and 7).

Theoretical analysis was harder to do than any other technique except the replicated project.



4.3 Cultu

By compar s in the
software er ing to our 7
criteria and o mostly of
professiona nt between
these two p 1.

Figure 5 su at the
validation c ment as the
most impor ernal validity
was of less crucial concern. That can be interpreted as the self-interest of industry in choosing methods applicable to its own environment and of less concern if it also aided a competitor.

On the other hand, for the research community of sample 1, internal and external validity, the ability of the validation to demonstrate effectiveness of the technique in the experimental sample and also to be able to generalize to other samples, were the primary criteria. Confirming a theory was next, obviously influenced

by the research community's orientation in developing new theoretical foundations for technology. At the other end of the scale, cost was of less concern where ease of replication was only 5th most important and cost of adding additional subjects was rated as last.

This points out some of the problems we addressed at the beginning of this paper. The research community is more concerned with theory confirmation and validity of the experiment and less concerned about costs, whereas the industrial community is more concerned about costs and applicability in their own environment and less concerned about general scientific results which can aid the community at large.

4.4 Composite measures

Given the set of 7 criteria, can we generate any composite measure for evaluating the effectiveness of the various validation methods? Since we have the respondents' impressions of the importance of each of the 7 criteria (via Figure 5), one obvious composite measure is the weighted sum of all the criteria evaluations. In this case, low score would determine the most significant methods. Table 4.1 gives these results.

Table 4.1 Composite measures

Sample 1 ordering (Research group)		**Sample 3 ordering (Student group)**		**Sample 2 ordering (Industry group)**	
Simulation	288	Case study	284	Measurement	258
Static analysis	292	Legacy data	314	Data mining	305
Dynamic analysis	298	Field study	315	Theoretical analysis	324
Project monitoring	301	Simulation	333	Research literature	325
Lessons learned	339	Dynamic analysis	355	Case study	326
Legacy data	345	Static analysis	361	Field study	327
Synthetic study	346	Literature search	370	Pilot study	329
Theoretical analysis	348	Replicated experiment	387	Feature benchmark	338
Field study	363	Project monitoring	388	Survey	343
Literature search	367	Lessons learned	391	Demonstrator project	345
Replicated experiment	368	Theoretical analysis	405	Replicated project	361
Case study	398	Synthetic study	418	State of the art	407
				Vendor opinion	469

Table 4.1 reveals some interesting observations:

1. For the research community, tools-based techniques dominate the rankings. Simulation, static analysis, and dynamic analysis are techniques that are easy to automate and can be handled in the laboratory. On the other hand, techniques that are labor intensive and require interacting with industrial groups (e.g., replicated experiment and case study) are at the bottom of the list. From our own anecdotal experiences over the past 20 years, working with industry on real projects certainly is harder to manage than building evaluation tools in the lab.
2. For the industrial community (the student sample 3 population), almost the opposite seems true. Those techniques that can confirm a technique in the field using industry data (e.g., case study, legacy data, field study) dominate the rankings, while "artificial" environments (e.g., theoretical analysis, synthetic study) are at the bottom. Again, this seems to support the concept that industrial professionals are more concerned with effectiveness of the techniques in live situations than simply validating a concept.

3. The industrial group evaluating the industrial validation methods (sample 2) cannot be compared with the above two groups since the methods they evaluated were different; however, there are some interesting observations. For one, measurement, the continual collection of data on development practices, clearly dominates the ranking. This is a surprising considering the difficulty the software engineering measurement community has been having in getting industry to recognize the need to measure development practices. With models like the Software Engineering Institute's Capability Maturity Model (CMM), the SEI's Personal Software Process (PSP) and Basili's Experience Factory promoting measurement, perhaps the word is finally getting out about the need to measure. But actual practice does not seem to agree with the desires of the professionals in the field. In addition, theoretical analysis came out fairly high in this composite score, but that does not seem to relate to experiences in the field.
4. Also within the industrial group, the need to be state of the art came near the bottom of the list (12^{th} out of 13) as not important. Basing decisions on vendor opinions was last. Yet vendors often influence the decision making process. Vendor opinions were judged to be least effective with respect to internal and external validity (Figure 3), but since vendor opinion was also judged to be one of the easiest to do, apparently users rely on such opinions even though they know the results are not to be trusted.
5. Data mining of collected data turned out to be second most important according to the industrial group. This is compatible with measurement being most important. If data is not collected, then there is nothing available to mine. Theoretical validation, literature search, and various experimental developments (i.e., field study, case study, pilot study) all ranked about the same level of importance to this group.

5. Conclusions

In this paper we discuss a survey taken from approximately 90 software engineering professionals. The survey evaluated subjective opinions on the value of validation methods for transferring new technology into industry. The idea was to study those methods used by the research community to validate new technologies and those methods used by industry to evaluate a new technology and to try and understand the differences. From this survey, we can make the observation that the research community and the development community do indeed have different perceptions of the role of experimentation to validating new technology. Researchers are more interested in how well a theory has been validated, whereas industry is more attuned, as expected, to how well the technique works in their own environment. Costs, while important to the industry sample, are mostly ignored by the research community.

Publication of research results is a major focus of the research community. In this respect, journal editors can play an important role in affecting this cultural difference. Developing new technologies and getting them into use should be a major focus of software engineering research. Editors of journals consider requiring more real-world validation using models like case studies, legacy data and field studies and be more suspect at validation via laboratory models, such as simulation and synthetic studies.

The survey also indicates that one should not simply be state of the art simply to be "fashionable" or listen to vendors for technology transfer decisions. Such decisions should depend on more technological reasons. Yet such actions are taken daily.

Measurement became the most important industrial decision making process in our composite analysis, yet anecdotal evidence indicates that much of industry does not collect the necessary data to build measurement programs. For the most part, our earlier survey [Zelkowitz98], the composite scores, and the results in

Tables 3.1 to 3.3 are compatible. In the earlier survey, papers studied from 1995 used case study and lessons learned equally, followed by simulation at half that number. In Table 3.3, the student population considering the research techniques ranked case study as practical in six of the seven questions. The industrial group (Table 3.2) selected either measurement or case study as practical for six of the seven questions, but the researchers find case study either impractical or neutral. Case study requires collection of data and measurement. It appears that the industry population values these measurement techniques as important, cost is a significant driver to industry, measurement techniques are perceived as too expensive. Better methods and tools for aiding measurement techniques are required to address industry concerns and to make the techniques more acceptable to researchers.

Given that industry is most concerned with internal validity, better tools are needed to aid the research community so that external validity can be conveyed more effectively to the industrial community. This would limit the effects of the "silver bullet" solution to complex problems. Studies are needed to identify:

1. What are the primary drivers that affect applicability in different environments?
2. How do you measure the effectiveness of a new method in a different environment?

Some of the results obtained here may be viewed as obvious, but we believe that these opinions have not been quantified previously. The industrial and the research community do look at method validation for different purposes, so it is not too surprising that one does not share the beliefs of the other. This leads to conflicts when one group does not provide or use the results of the other.

Given the set of techniques described here, it would aid both communities if those techniques near the top of the rankings had better tool support. Measurement is clearly important to the industrial professional, so less expensive data collection methods are needed. Tools for collecting defect data or analyzing defect and resource data are needed. Tools to better evaluate case studies would help. How to deal with the high cost and poor perception of the replicated experiment needs to be further studied.

In this paper, as with our earlier survey of the research literature, we have tried to understand the process that organizations use to evaluate new technologies and transition them into industrial use. We haven't solved the significant technology transition problems with this survey, but we do believe we have indicated where further research is needed and why some of the current problems in technology transition exist. We need to further understand both cultures in order to determine which technique can best enable industry to make intelligent choices on which new technology to use and, we emphasize the need for research to develop the methods and tools to make these techniques practical..

Acknowledgments

We thank Dr. Nien Zhang for his suggestions regarding statistical methods for viewing this data.

References

[Brown96] Brown A. W. and K. C. Wallnau, A framework for evaluating software technology, *IEEE Software*, (September, 1996) 39-49.

[Fenton94] Fenton N., S. L. Pfleeger, and R. L. Glass, Science and substance: A challenge to software engineers, *IEEE Software*, Vol. 11, No. 4, 1994, 86-95.

[Daly97] Daly, J., K. El Emam, and J. Miller, Multi-method research in software engineering, 1997 IEEE Workshop on Empirical Studies of Software Maintenance (WESS '97) Bari, Italy, October 3, 1997.

[Tichy95] Tichy W. F., P. Lukowicz, L. Prechelt, and E. A. Heinz, Experimental evaluation in computer science: A quantitative study, *J. of Systems and Software* Vol. 28, No. 1, 1995 9-18.

[Tichy98] Tichy, W., Should computer scientists experiment more?, *Computer*, Vol.31, No.5, 1998, pp. 32-40.

[Zelkowitz97] Zelkowitz M. and D. Wallace, Experimental validation in software engineering, *Information and Software Technology*, Vol. 39, 1997, 735-743.

[Zelkowitz98] Zelkowitz M. and D. Wallace, Experimental models for validating technology, *Computer*, Vol.31, No.5, 1998, 23-31.

APPENDIX 1 -- Types of Research Validation

1. **Case study** - a project is monitored and data collected over time. Data collection is derived from a specific goal for the project. A certain attribute is monitored (e.g., reliability, cost) and data is collected to measure that attribute.

2. **Dynamic analysis** - a product is executed for performance. Many methods instrument the given product by adding debugging or testing code in such a way that features of the product can be demonstrated and evaluated when the product is executed.

3. **Legacy data** - data from previous projects is examined for understanding in order to apply that information on a new project under development. Available data includes all artifacts involved in the product, e.g., the source program, specification, design, and testing documentation, as well as data collected in its development.

4. **Lessons-learned** - qualitative data from completed projects is examined. Lessons-learned documents are often produced after a large industrial project is completed. A study of these documents often reveals qualitative aspects which can be used to improve future developments.

5. **Literature search** - previously published studies are examined. It requires the investigator to analyze the results of papers and other documents that are publicly available (e.g., conference and journal articles).

6. **Project monitoring** - collect and store development data during project development. The available data will be whatever the project generates with no attempt to influence or redirect the development process or methods that are being used.

7. **Field study** - A field study may examine data collected from several projects (e.g., subjects) simultaneously. Typically, data are collected from each activity in order to determine the effectiveness of that activity. Often an outside group will monitor the actions of each subject group, whereas in the case study model, the subjects themselves perform the data collection activities.

8. **Replicated experiment** - develop multiple versions of product. In a replicated experiment several projects are staffed to perform a task in multiple ways. Control variables are set (e.g., duration, staff level, methods used) and statistical validity can be more applied. This is the "classical" scientific experiment where similar process is altered repeatedly to see the effects of that change.

9. **Simulation** - execute product with artificial data. Related to dynamic analysis is the concept of simulation. We can evaluate a technology by executing the product using a model of the real environment. We hypothesize, or predict, how the real environment will react to the new technology.

10. **Static analysis** - examine structure of developed product. This is a special case of studying legacy data except that we centralize our concerns on the product that was developed, whereas legacy data also included development process measurement.

11. **Synthetic environment** - replicate one factor in laboratory setting. In software development, projects are usually large and the staffing of multiple projects (e.g., the replicated experiment) in a realistic

setting is usually prohibitively expensive. For this reason, most software engineering replications are performed in a smaller artificial setting, which only approximates the environment of the larger projects.

12. **Theoretical analysis** - uses logic to validate a theory; validation consists of logical proofs derived from a specific set of axioms.

APPENDIX 2 -- Types of Industrial Evaluation

1. **Case study** -- Sample projects, typical of other industrial developments for that organization, are developed, where some new technology is applied and the results of using that technology are observed.

2. **Data mining** -- Completed projects are studied in order to find new information about the technologies to develop those projects.

3. **Demonstrator projects** -- Multiple instances of an application, with essential features deleted, are built in order to observe behavior of the new system.

4. **Feature benchmark** -- Alternative technologies are evaluated and comparable data are collected. This is usually a "desk study" using documentation on those features.

5. **Field study** -- An assessment is made by observing the behavior of several other development groups over a relatively short time.

6. **Measurement** -- Data is continually collected on development practices. This data can be investigated when a new technology is proposed.

7. **Pilot study** - A sample project that uses a new technology. This is generally a smaller application (than a case study) before scaling up to full deployment, but is more complete than a demonstration project.

8. **Research literature** -- Information is obtained from professional conferences, journals, and other academic sources of information.

9. **Shadow (Replicated) project** -- One or more projects duplicate another project in order to test different alternative technologies on the same application.

10. **State of the art** -- Using a new technology that is based upon purchaser or client desires or government rules to only use the latest or best technology.

11. **Survey** -- Experts in other areas (e.g., other companies, academia, other projects) are queried for their expert opinion of the probable effects of some new technology.

12. **Theoretical analysis** -- Basing an opinion on the validity of the mathematical model of a new technology.

13. **Vendor opinion** -- Vendors (e.g, through trade shows, trade press, advertising, sales meetings) promote a new technology.